

Effect Heterogeneity and Variable Selection for Standardizing Experimental Findings

Anders Huitfeldt¹, Sonja A. Swanson^{2,3}, Mats J. Stensrud⁴ and Etsuji Suzuki⁵

¹The Meta-Research Innovation Center at Stanford, Stanford University

²Department of Epidemiology, Erasmus MC

³Department of Epidemiology, Harvard T.H. Chan School of Public Health

⁴Department of Biostatistics, University of Oslo

⁵Department of Epidemiology, Okayama University

October 4, 2016

Abstract

Randomized trials are rarely representative of the general population. To account for this, results may be standardized to a target population which may be different from the one in which the trial was conducted. In this paper, we discuss three different approaches for reasoning about which covariates must be standardized over. This includes the traditional approach based on standard effect measures such as the risk difference or the risk ratio, Bareinboim and Pearl's graphical approach based on selection diagrams, and a new approach based on Counterfactual Outcome State Transition parameters. We compare and contrast these approaches, and argue that the choice must be informed by expert beliefs about which assumptions are most likely to be approximately true in the specific scientific context.

1. BACKGROUND

The participants in randomized trials are often not representative of the populations in which the results will be used to inform clinical decisions. Several different statistical methods have been proposed to standardize experimental results over some set of baseline covariates, in order to account for differences between the study population and the target population. However, less attention has been given to how an investigator should reason

about which covariates need to be standardized for. This choice of variables is important not only for the standardization procedure itself, but also for determining which personal characteristics must be considered when quantifying how representative the trials are. In this paper, we discuss different strategies for selecting such covariates, and show that these strategies are associated with different approaches to operationalizing effect homogeneity.

Epidemiologists have traditionally defined effect homogeneity in terms of effect measures that compare the distribution of the counterfactual outcome under an active treatment with the distribution of the counterfactual outcome under an alternative. Examples of this include the risk ratio, the risk difference and the odds ratio. These definitions of effect homogeneity are associated with several well-known conceptual shortcomings, including lack of biological interpretation, baseline risk dependence, zero bounds, prediction outside the range of valid probabilities, non-collapsibility and asymmetry [1]

There have also been several recent methodological developments in defining effect homogeneity based on counterfactual distributions rather than specific measures of effect. Specifically, these approaches define effect homogeneity in terms of individual counterfactual distributions such as the outcome under the active treatment or (separately) the outcome under standard of care. VanderWeele described this approach as “Effect modification in distribution” [2]. Although methods based on assuming conditional homogeneity of the distribution of a counterfactual variable are mathematically elegant and avoid most of the limitations of defining homogeneity with respect to effect measures, they make strong assumptions which will often be violated in realistic epidemiologic settings. Table 1 shows an overview of different ways an investigator can operationalize effect homogeneity.

The paper is organized as follows. First, we review approaches to variable selection that are based on conditional homogeneity of standard effects measures such as the risk ratio and the risk difference, and discuss some of the shortcomings that have limited the generality of these approaches. We then review approaches to variable selection based on conditional homogeneity of individual counterfactual distributions, with a particular emphasis on methods based on causal diagrams for research transportability [3]. Finally, we propose an approach to variable selection based on the recently introduced counterfactual outcome state transition (COST) parameters, which overcomes many of the shortcomings of traditional effect measures [1]. This approach makes different assumptions than those embedded in approaches based on conditional homogeneity in distribution such as Bareinboim and Pearl's selection diagrams.

For all examples, we will consider the effect of binary treatment A (for example, a drug) on binary outcome Y (for example, a side effect). We will use the letter V to refer to baseline covariates which are potential effect modifiers (for example, gender or nationality). Counterfactuals will be denoted using superscripts [4]. For example, $Y^{a=0}$ is an indicator for whether an individual would have got side effect Y if, possibly contrary to fact, she did not initiate treatment with drug A . We will consider several measures of causal effect including the Causal Risk Difference ($RD = \Pr(Y^{a=1} = 1) - \Pr(Y^{a=0} = 1)$), the Causal Risk

Table 1: Definitions of Conditional Effect Homogeneity

	Definition
Effect Homogeneity in Measure	
On the Risk Difference Scale	$RD_s = RD_t$ (Conditional on $V = v$)
On the Risk Ratio Scale	$RR_s = RR_t$ (Conditional on $V = v$)
On the Odds Ratio Scale	$OR_s = OR_t$ (Conditional on $V = v$)
Effect Homogeneity in Distribution	
S-ignorability	$Y^a \perp\!\!\!\perp P V$ for all values of a
S-admissibility	$Y^a \perp\!\!\!\perp P^a V^a$ for all values of a
Homogeneity of COST Parameters	
For Introducing Treatment	$Y^{a=1} \perp\!\!\!\perp P Y^{a=0}, V = v$
For Removing Treatment	$Y^{a=0} \perp\!\!\!\perp P Y^{a=1}, V = v$

Ratio ($RR(-) = \frac{\Pr(Y^{a=1}=1)}{\Pr(Y^{a=0}=1)}$), and the Causal Odds Ratio ($OR = \frac{\frac{\Pr(Y^{a=1}=1)}{1-\Pr(Y^{a=1}=1)}}{\frac{\Pr(Y^{a=0}=1)}{1-\Pr(Y^{a=0}=1)}}$). These effect measures may be defined in a specific population, which we denote using superscript. For instance, RD_t is the risk difference in population p .

We will consider two closely related classes of problems: In problem type 1, which arises commonly in the context of external validity and research generalization, we will assume that we have experimental evidence for the causal effect of the drug in study population $P = s$ and that we wish to predict the effect of introducing the drug in the target population $P = t$, in which we only have observational data ("Transportability") [5]. In problem type 2, which arises commonly in meta-analysis and model specification, we have experimental data from two or more populations that we wish to summarize under a plausible homogeneity assumption ("Data Fusion"). [6]

2. EFFECT HOMOGENEITY IN MEASURE

Effect homogeneity in measure occurs whenever the effect in one population is equal to the effect in another population in terms of a particular effect measure, such as the risk difference or the risk ratio. For example, if the risk difference in population $P = s$ (i.e. RD_s) is equal to risk difference in population $P = t$ (i.e. RD_t) we say that there is effect homogeneity on the risk difference scale. If this condition holds within levels of a set of measured covariates V we say that there is conditional effect homogeneity on that scale, and that V is a sufficient set of effect measure modifiers for the transportation from s to t .

To illustrate, it is possible that the risk ratio for adverse effects of Codeine differs between Norway and Japan because the two countries have different distributions of variants of CYP2D6 [7], a gene associated with drug metabolism, but that the risk ratio associated

with the use of the drug is equal between Norwegians and Japanese who have the same variant of the gene. If that is the case, then we have effect homogeneity conditional on CYP2D6 variant, and CYP2D6 is a sufficient set of effect modifiers. It must be pointed out that a sufficient set of effect measure modifiers may not exist.

Many commonly used methods rely on assumptions that are equivalent to conditional effect homogeneity in measure. For example, in the data fusion setting, the logistic regression model

$$\text{logit Pr}\{(Y = 1|A, P, V)\} = \beta_0 + \beta_1 A + \beta_2 P + \beta_4 V\}$$

which omits a product term $\beta_3 \times A \times P$, makes the assumption that the odds ratio of A on Y in the group $P = s$ is equal to the odds ratio in $P = t$, conditional on $V = v$, or in other words, that there is conditional effect homogeneity on the odds ratio scale.

Another example of a method that relies on conditional effect homogeneity in measure, this time in the setting of transportability, occurs when an investigator attempts to account for heterogeneity between populations by standardizing an effect measure over a set of covariates V . In Table 2, we show simple formulas for standardizing experimental results to a target population over a sufficient set of effect modifiers. Approach one, which is a weighted average of the effect measure, is valid for collapsible effect measures [8], whereas approach two, which is a weighted-average of the stratum-specific predicted risks under treatment, is valid for any effect measure.

These methods often have to be justified on the basis of background expert knowledge. This raises the question of how an investigator would be able to credibly invoke *a priori* knowledge that V is a sufficient set of effect modifiers on a particular effect scale. In general, convincing arguments for modeling assumptions will take the form of an explicitly stated data generating mechanism, i.e. a clearly outlined chain of events which guarantees that the necessary conditions will hold. To our knowledge, no non-parametric mechanism has been proposed that would guarantee conditional effect homogeneity on either the risk difference, risk ratio or odds ratio scale, without also guaranteeing effect homogeneity in distribution, which is discussed later.

3. EFFECT HOMOGENEITY IN DISTRIBUTION

An alternative approach is to operationalize effect homogeneity in terms of the individual counterfactual distributions under treatment and no treatment. Effect homogeneity in distribution holds whenever the following two conditions hold simultaneously: (1) If everyone in both populations were untreated, you would observe the same distribution of outcomes in the two populations ($Y^{a=0} \perp\!\!\!\perp P$) and (2) if everyone in both populations were treated, you would observe the same distribution of outcomes in the two populations ($Y^{a=1} \perp\!\!\!\perp P$). VanderWeele [2] proved that effect homogeneity in distribution implies effect homogeneity in

Table 2: Three Approaches to Standardization

	Measure Standardized	Being Standardized Over	Covariates Standardized Over	Validity Conditions
$RR_t = \sum_v RR_{s,v} \times w_{v,t}$	The Effect Measure		A Sufficient Set of Effect Modifiers	Conditional Effect Homogeneity in Measure and Collapsible Effect Measure
$\Pr(Y^{a=1} P = t) = \sum_v \Pr(Y^{a=0} P = t, V = v) \times RR_{s,v} \times \Pr(V = v P = t)$	The Predicted Average Outcome under a Specific Effect Measure		A Sufficient set of Effect Modifiers	Conditional Effect Homogeneity in Measure
$\Pr(Y^a P = t) = \sum_v \Pr(Y^a V = v, P = s) \times \Pr(V = v P = t)$	The Average Outcome		Variables sufficient to block all pathways between P and Y	Conditional Effect Homogeneity in Distribution

measure for all standard effect measures, effect homogeneity in distribution is therefore a strictly stronger assumption than effect homogeneity in measure.

As with effect homogeneity in measure, this condition may hold within levels of a set of covariates V . If effect homogeneity in distribution holds conditional on V , one can use a third standardization formula, also shown in Table 2, based on separately standardizing the risk under treatment and the risk under no treatment from the study population, to the distribution of V in the target population. This standardization formula can equivalently be computed by using Cole and Stuart's inverse probability weighted methods [9].

Effect homogeneity in distribution will often be implausible in realistic randomized and observational studies. Specifically, methods that rely on this assumption are valid only if they account for every cause of the outcome that differs between the study population and the target population.

Moreover, these approaches do not make use of possible information contained in the joint counterfactual distributions that is not contained in the individual counterfactual distributions. To illustrate, consider a situation where we have conducted a randomized controlled trial on the effect of placebo vs no treatment on the incidence of cardiovascular disease, and concluded that the effect in the study population is null. We are interested in predicting the effect in a different target population, but we believe there may be unmeasured causes of cardiovascular disease that differ between the study population and the target population. In such situations, if we use a notion of effect homogeneity in distribution, we are likely forced to conclude that we are unable to make predictions for the target population. In contrast, investigators using an approach based on effect homogeneity in measure may be able to clarify plausible conditions under which the summary measure can be extrapolated to the target population.

In the previous section, we discussed the logistic regression model

$$\text{logit Pr}\{(Y = 1|A, P, V)\} = \beta_0 + \beta_1 A + \beta_2 P + \beta_4 V\}$$

which omits the product term $\beta_3 \times A \times P$, and showed that this model is justified under conditional effect homogeneity in measure on the odds ratio scale. We note that this model could also be justified under conditional effect homogeneity in distribution. However, this modeling approach has a surprising implication: If effect homogeneity in distribution holds and A is unconfounded, then β_2 must be equal to zero. This makes the model subject to an empirical test: if e.g. the Wald test rejects $\beta_2 = 0$, the model is misspecified. We hold that this suggests that effect homogeneity in distribution is a very strong concept, and that investigators often have to rely on a weaker form of effect homogeneity.

4. SELECTION DIAGRAMS

One example of a data generating mechanism that guarantees effect homogeneity in distribution (and therefore also effect homogeneity in measure for all possible effect measures)

was provided by Bareinboim and Pearl, based on causal diagrams. These diagrams are, to our knowledge, the only published formal framework for reasoning about which variables to adjust for when using approaches based on effect homogeneity in distribution.

A selection diagram is constructed as follows: First, the investigator must provide a causal directed acyclic graph (DAG) that is valid both for the study population and for the target population. For this to be possible, the variables must be in the same temporal order between the two populations. If that requirement is met, a DAG which is valid for both populations can be constructed by including every node and edge from the causal DAG in each population. After a shared causal DAG has been constructed, one must also add (1) a selection variable node P , (2) All causes of the outcome whose distributions differ between the populations $P = s$ and $P = t$ (including those causes that act through intermediates on the graph), and (3) all paths between P and Y that one is not able to rule out based on the temporal structure or expert knowledge.

Once a selection diagram has been constructed, one can check for transportability of the results by determining whether Y is d-separated from P , given some set of measured variables V . If such d-separation holds, there will exist a transport formula which identifies the causal effect in the target population based on a combination of observed quantities in the study population and observed quantities in the target population. If V consists only of baseline covariates, then the transport formula is equal to the standardization formula discussed in the previous section.

The independence relation that is queried by this d-separation approach can be written algebraically as

$$Y^a \perp\!\!\!\perp P^a | V^a$$

which Bareinboim and Pearl referred to as “S-admissibility”. When P and V are pre-treatment variables, $P^a = P$ and $V^a = V$, so the independence relation can be simplified as

$$Y^a \perp\!\!\!\perp P | V$$

(or “S-ignorability”), or equivalently as $f(Y^a | P=s, V=v) = f(Y^a | P=t, V=v)$ for all values of a . This is identical to the definition of conditional effect homogeneity in distribution, which illustrates the equivalence between the graphical approach and approaches based on effect homogeneity in distribution, when V and P are pre-treatment.

We next proceed to show that in many transportation problems of interest in clinical medicine and epidemiology, adjustment for post-baseline covariates is not necessary. To do so, we will discuss certain conditions which greatly reduce the complexity of the problem, and which are reasonable approximations in many applications including most meta-analyses of clinical trials.

(1) Selection into the trial is not a downstream consequence of treatment assignment. Therefore, $P = s$ implies $P^a = s$ for all values of a . In most randomized controlled trials, this condition is expected to hold by design, though we note that there are exceptions including

Zelen's design [10]. If there are only two values of P this further implies that $P = P^a$ for all individuals.

(2) Treatment is unavailable outside of the randomized controlled trial. Therefore, $P = t$ implies $P^{a=0} = t$ by consistency. (This condition is only necessary if P can take more than two values, i.e. if some people are neither in the study population nor in the target population)

(3) If there exists a post-baseline covariate which is a marker for selection into the trial, then there also exists some measured baseline covariate which is a better marker for the selection process (for example because the investigators used the baseline covariate when they recruited the study participants). In other words, $Y^a \perp\!\!\!\perp P|V^a$ implies the existence of some V such that $Y^a \perp\!\!\!\perp P|V$

When these three conditions hold, s-admissibility is equivalent to s-ignorability, and an investigator with experimental or otherwise unconfounded data will therefore not have to worry about post-randomization adjustment for transportability. Note that while it is certainly possible that there is selection out of the trial based on post-baseline variables, this is best considered as a form of selection bias [11] rather than a transportability problem. In other words, people who drop out of the trial should still be considered part of the population, and appropriate methods should be used to adjust for the selection bias which results from having missing data on these individuals.

5. COUNTERFACTUAL OUTCOME STATE TRANSITION PARAMETERS

The COST parameters G and H are effect measures for binary outcomes which have important theoretical advantages over traditional effect measures such as the risk ratio or odds ratio. Briefly, we recall these parameters are defined as follows:

$$G = \Pr(Y^{a=1} = 1 \mid Y^{a=0} = 1)$$

$$H = \Pr(Y^{a=1} = 0 \mid Y^{a=0} = 0)$$

In Huitfeldt et al (working paper 2016) [1], the effect of introducing treatment was defined to be equal between the two populations to be equal if and only if $G_s = G_t$ and $H_s = H_t$, which can equivalently be written as $Y^{a=0} \perp\!\!\!\perp P|Y^{a=1}$. Such effect equality may hold within levels of measured covariates V , in which case, there is conditional equality of the effect of introducing or removing treatment.

One key advantage of this definition is that it correspond to a biological interpretation: In a population where everyone is untreated, the COST parameters can be interpreted as the proportion of cases and non-cases that would not have had the opposite outcome if their exposure status had been altered. These proportions may be equal between similar groups in different populations, for instance because response to treatment is associated with a genotype which has equal prevalence between the corresponding subgroups in the

two populations. Under this definition of effect equality, it is therefore sufficient to account for all variables which are associated with treatment response, rather than all variables that are causes of the outcome.

In contrast to standard effect measures, effect homogeneity in distribution ($Y^a \perp\!\!\!\perp P$ for all values of a) is not sufficient to guarantee homogeneity of the COST parameters, unless the condition is slightly strengthened such that the counterfactual distributions are *jointly* independent of P , i.e. $(Y^{a=0}, Y^{a=1}) \perp\!\!\!\perp P$. Note that in any model based on selection diagrams in which the individual independences hold, this joint independence will also hold.

If the COST parameters are equal between the study population and the target population conditional on covariates V , the results from the trial may be standardized to the target population using either approach 1 or approach 2 from Table 1. Using the first of these, the effect measure G can be standardized using the weights $w_v = \Pr(V = v | Y^{a=0} = 1)$, whereas the effect measure H can be standardized using the weights $w_v = \Pr(V = v | Y^{a=0} = 0)$. Using the second approach, the stratum-specific values of $\Pr(Y^{a=1} = 1 | V = v, P = t)$ can be computed separately from the stratum-specific effect measures and $\Pr(Y^{a=0} = 1 | V = v, P = t)$, and standardized to the target population using the weights $\Pr(V = v)$.

Because the COST parameters are only identified under an assumption of monotonicity, these methods should be used carefully. The bias which is associated with the use of COST parameters is small either if non-monotonicity is negligible, or if the baseline risks are similar between the target population and the study population. We note that the closer one gets to conditioning on all causes of the outcome, the more similar the baseline risks will be, which means that including additional variables can sometimes reduce the remaining bias associated with non-monotonicity.

Another important limitation of COST parameters is that they have so far only been defined only for binary outcomes. Extensions to continuous outcomes and survival data are not trivial, and it remains to be seen whether such extension will be feasible.

6. CONCLUSIONS

Effects often differ between populations, and investigators will often have to standardize experimental results over a set of effect modifiers. Before it is possible to begin reasoning about which covariates must be standardized over, it is necessary to provide a definition of effect homogeneity. Several different approaches have been proposed.

The approach based on COST parameters requires that the investigators have accounted for all variables that predict treatment response, that only baseline covariates are necessary for this purpose, and that the effects of treatment are monotonic. In contrast, the approach based on selection diagrams does not require monotonicity and generalizes to selection processes that are downstream from treatment, but this comes at the cost of requiring the investigators to account for all covariates that differ between the two populations and which are correlated with the counterfactual outcome

The choice between the two approaches will therefore depend on expert beliefs about which assumptions are most likely to be approximately true in the specific scientific context. We believe that at least in some situations, the assumptions underlying the approach based on COST parameters are less restrictive than the assumptions underlying selection diagrams and other approaches based on effect homogeneity in distribution.

Finally, we note that investigators who use methods based on effect homogeneity in distribution will often report their conclusions in terms of a parametric effect measure. This suggests the possibility of using a hybrid approach, where one chooses a risk ratio model informed by COST parameters but standardizes over all causes of the outcome, not just those variables associated with treatment response. This approach will be valid either if the conditions of the COST parameter approach are met, or if there is S-ignorability conditional on V .

ACKNOWLEDGEMENT

The authors thank Steve Goodman and Miguel Hernan for discussions and helpful comments on earlier drafts of this manuscript.

AUTHOR CONTRIBUTIONS

AH had the original idea, provided the original version of the theorems and proofs, wrote the first draft of the manuscript and coordinated the research project. SAS, MJS and ES contributed original intellectual content and extensively restructured and revised the manuscript. All authors approved the final version of the manuscript.

CORRESPONDENCE

All correspondence should be directed to Anders Huitfeldt at The Meta-Research Innovation Center at Stanford, Stanford University School of Medicine, 1070 Arastradero Road, Palo Alto CA 94303; e-mail: ahuitfel@stanford.edu.

Anonymous feedback is welcomed at <http://www.admonymous.com/effectmeasurepaper>. Anders Huitfeldt invokes Crocker's Rules (<http://sl4.org/crocker.html>) on behalf of all authors for all anonymous and non-anonymous feedback on this manuscript.

REFERENCES

- [1] Anders Huitfeldt, Andrew Goldstein, and Sonja A. Swanson. The Choice of Effect Measure for Binary Outcomes: Introducing Counterfactual Outcome State Transition Parameters. *Unpublished Manuscript*, 2016.

- [2] Tyler J VanderWeele. Confounding and Effect Modification: Distribution and Measure. *Epidemiologic methods*, 1(1):55–82, 8 2012.
- [3] Elias Bareinboim and Judea Pearl. A General Algorithm for Deciding Transportability of Experimental Results. *Journal of Causal Inference*, 1(1):107–134, 1 2013.
- [4] Miguel A Hernán and James M Robins. *Causal Inference*. Chapman & Hall/CRC, Boca Raton, 2016.
- [5] Judea Pearl and Elias Bareinboim. External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science*, 29(4):579–595, 11 2014.
- [6] Elias Bareinboim and Judea Pearl. Causal Inference and the Data-Fusion Problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 7 2016.
- [7] Stephen Bernard, Kathleen A Neville, Anne T Nguyen, and David A Flockhart. Interethnic Differences in Genetic Polymorphisms of CYP2D6 in the U.S. Population: Clinical Implications. *The oncologist*, 11(2):126–35, 2 2006.
- [8] Anders Huitfeldt, Mats Julius Stensrud, and Etsuji Suzuki. On the Collapsibility of Effect Measures in the Counterfactual Causal Framework. *Working Paper*, 2016.
- [9] Stephen R Cole and Elizabeth A Stuart. Generalizing Evidence from Randomized Clinical Trials to Target Populations: The ACTG 320 Trial. *American journal of epidemiology*, 172(1):107–15, 7 2010.
- [10] Marvin Zelen. A New Design for Randomized Clinical Trials. *New England Journal of Medicine*, 300(22):1242–1245, 5 1979.
- [11] Miguel A Hernán, Sonia Hernández-Díaz, and James M Robins. A Structural Approach to Selection Bias. *Epidemiology (Cambridge, Mass.)*, 15(5):615–25, 9 2004.

A. APPENDIX 1

Here, we prove that if there is effect homogeneity in distribution, then the parameter β_2 must be equal to zero in the regression model

$$\text{logit Pr}\{(Y = 1|A, P, V)\} = \beta_0 + \beta_1 A + \beta_2 P + \beta_4 V \quad (1)$$

To do so, we will make the following assumptions:

$Y^a \perp\!\!\!\perp P|V$ for all values of a (Effect homogeneity in distribution)

$Y^a \perp\!\!\!\perp A|V, P$ for all values of a (Exchangeability)

$$Y^a = A \text{ if } A = a \text{ (Consistency)}$$

By consistency and exchangeability, the model can be rewritten as a structural model:

$$\text{logit } Pr(Y^a = 1|P, V) = \beta_0 + \beta_1 a + \beta_2 P + \beta_4 V \quad (2)$$

If $P = 0$, we have :

$$\text{logit } Pr(Y^a = 1|P, V) = \beta_0 + \beta_1 a + \beta_4 V \quad (3)$$

If $P = 1$ we have

$$\text{logit } Pr(Y^a = 1|P, V) = \beta_0 + \beta_1 a + \beta_2 P + \beta_4 V \quad (4)$$

By the assumption of effect homogeneity, we can set these equal:

$$\beta_0 + \beta_1 + \beta_4 = \beta_0 + \beta_1 + \beta_2 + \beta_4$$

Solving this for β_2 we get $\beta_2 = 0$